

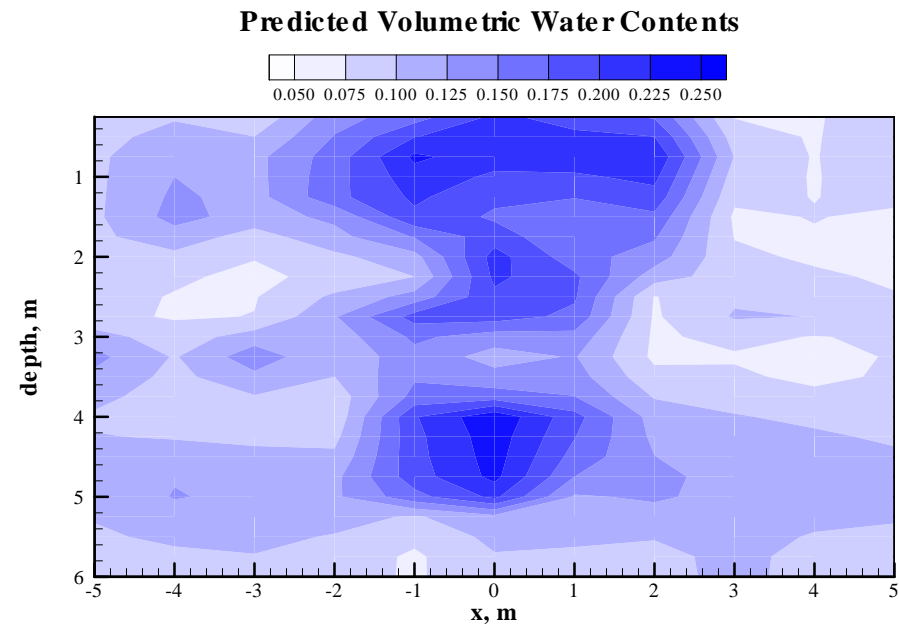
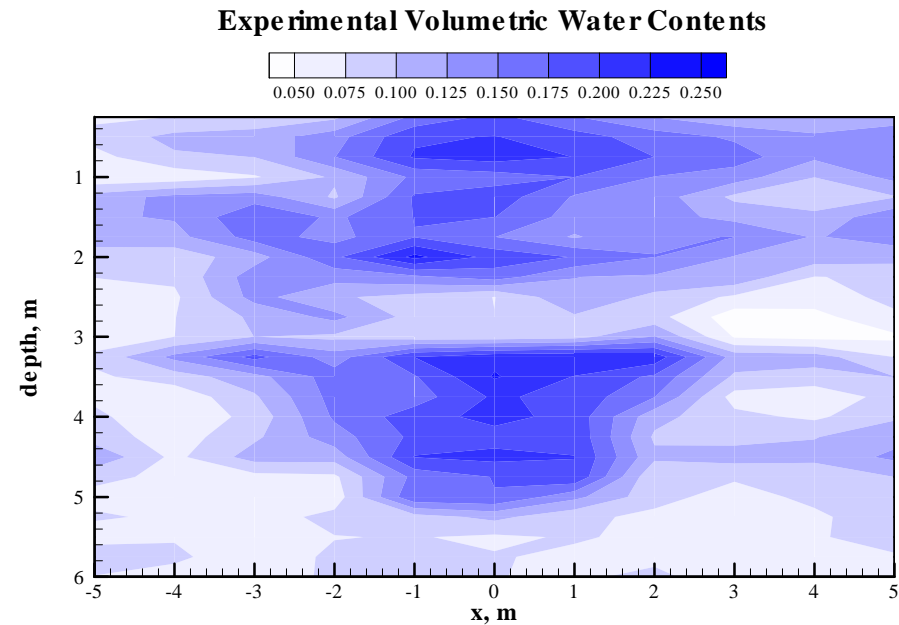
The Development of Model Validation Metrics using Probability Based Methods

Richard Hills
Mechanical Engineering
New Mexico State University
rhills@nmsu.edu

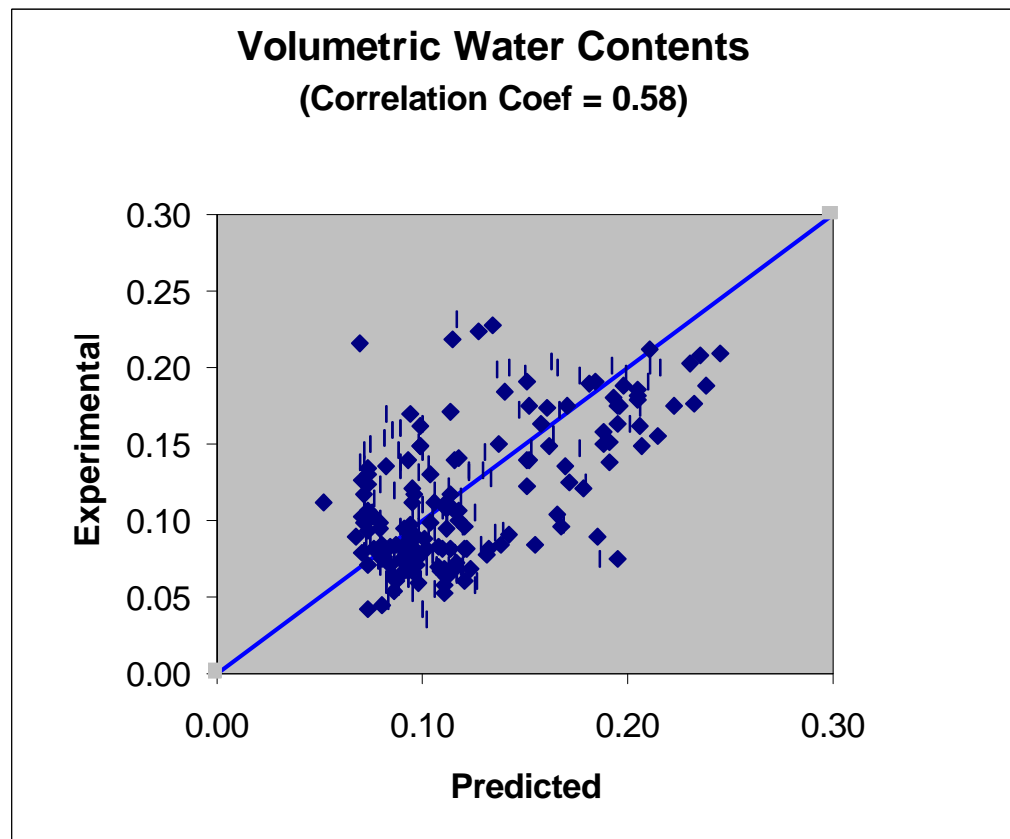
Supported by:
Sandia National Laboratories
Technical Contact: Tim Trucano

Is this model valid?

- Water plume data from the Las Cruces Trench Experiments
- Experiment: Water content measured on a 11 wide by 24 deep grid
- Characterization: Detailed characterization based on 450 soil samples taken in a vertical plane
- Prediction: One realization of a heterogeneous soil water retention model based on detailed characterization



How about this model?



Metrics

- How do we rigorously compare observations to predictions?
- Qualitatively
 - Graphical Comparisons
- Quantitative Metrics
 - Correlation coefficients, maximum error, mean error, etc.
- What values of these metrics are good enough?

Statistical Based Validation

- Use statistical methods to define acceptable values for a metric
- Scientific Validation - is difference between prediction and observation significant relative to the uncertainty in the validation exercise?
- Engineering Validation - is difference between prediction and observation significant relative to the uncertainty in the validation exercise plus some acceptable error?

Validation, continued

- Both questions require probabilistic models for the uncertainty to define whether the prediction errors are statistically significant relative to this uncertainty.
- Several approaches to develop probability models
 - Many independent, repeated experiments - usually not available for complex engineering applications
 - Fewer independent experiments, multivariate data
 - Evaluate the uncertainty using probabilistic model for model parameters and measurement uncertainty

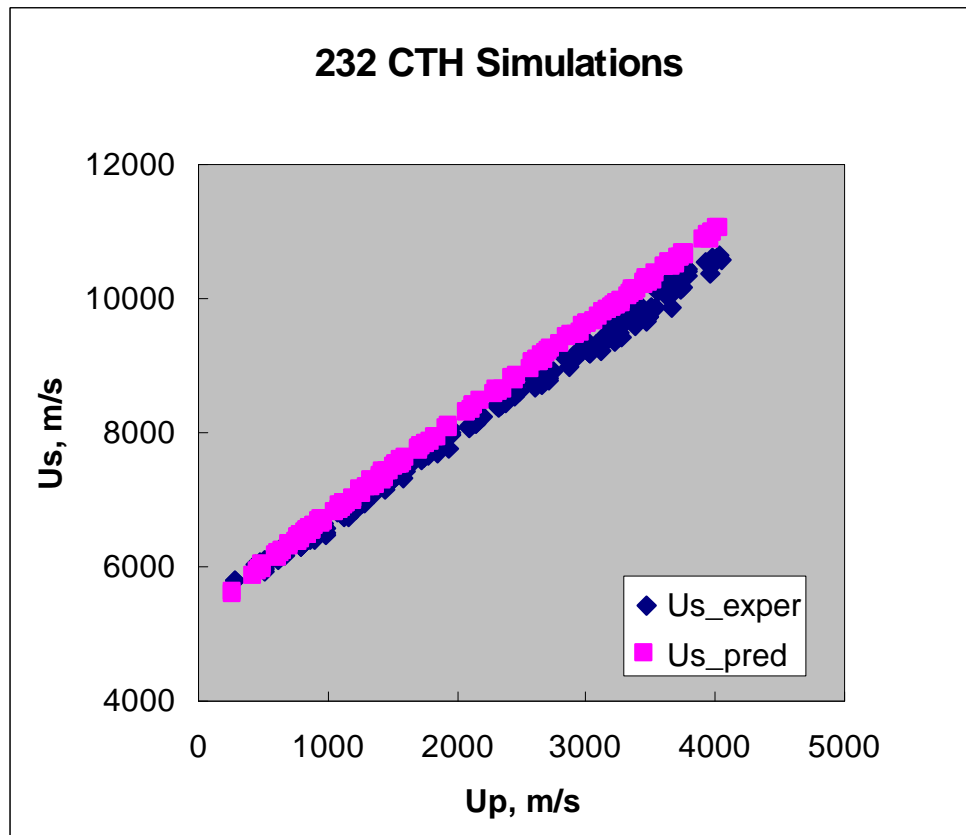
This Talk

- Background
- Standard statistical methods for validation using data from multiple experiments - CTH example
- Validation using multivariate data using propagation of uncertainty analysis - Conceptual

CTH Example

- Model :
 - 1-D CTH simulation (based on the 1-D example provided with CTH)
 - Impact aluminum slug with velocity of $2*U_p$ on a stationary aluminum slug of equal size - results in shock with the behind the shock particle speed of U_p
 - EOS - Sesame 2024 aluminum
 - Evaluate corresponding shock wave speed U_s
- Experimental Data:
 - U_s vs. U_p data obtained from LANL Shock Wave Compendium
 - 232 experiments - presumed independent

CTH Results

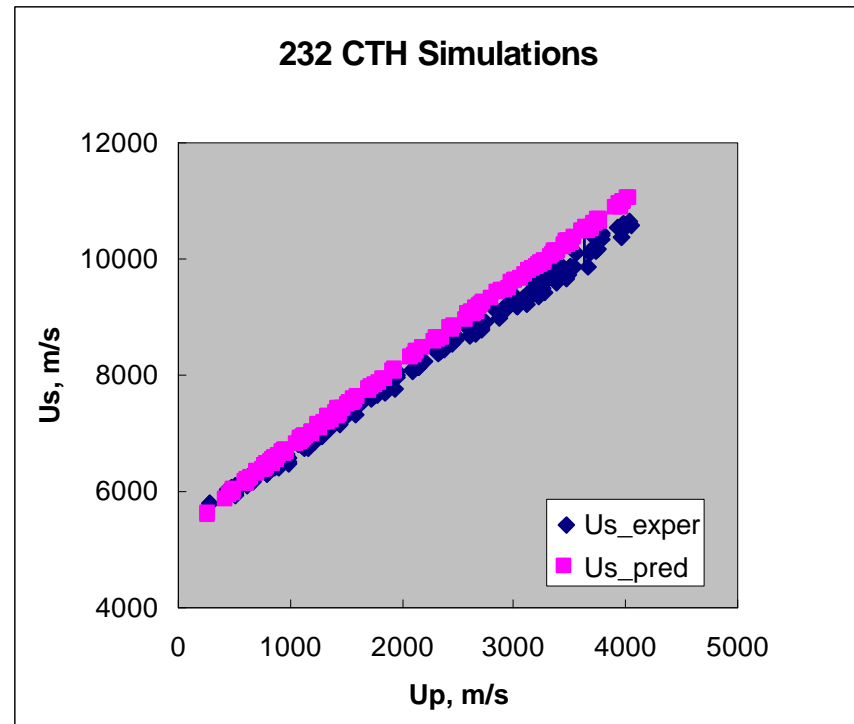


Is this model valid?

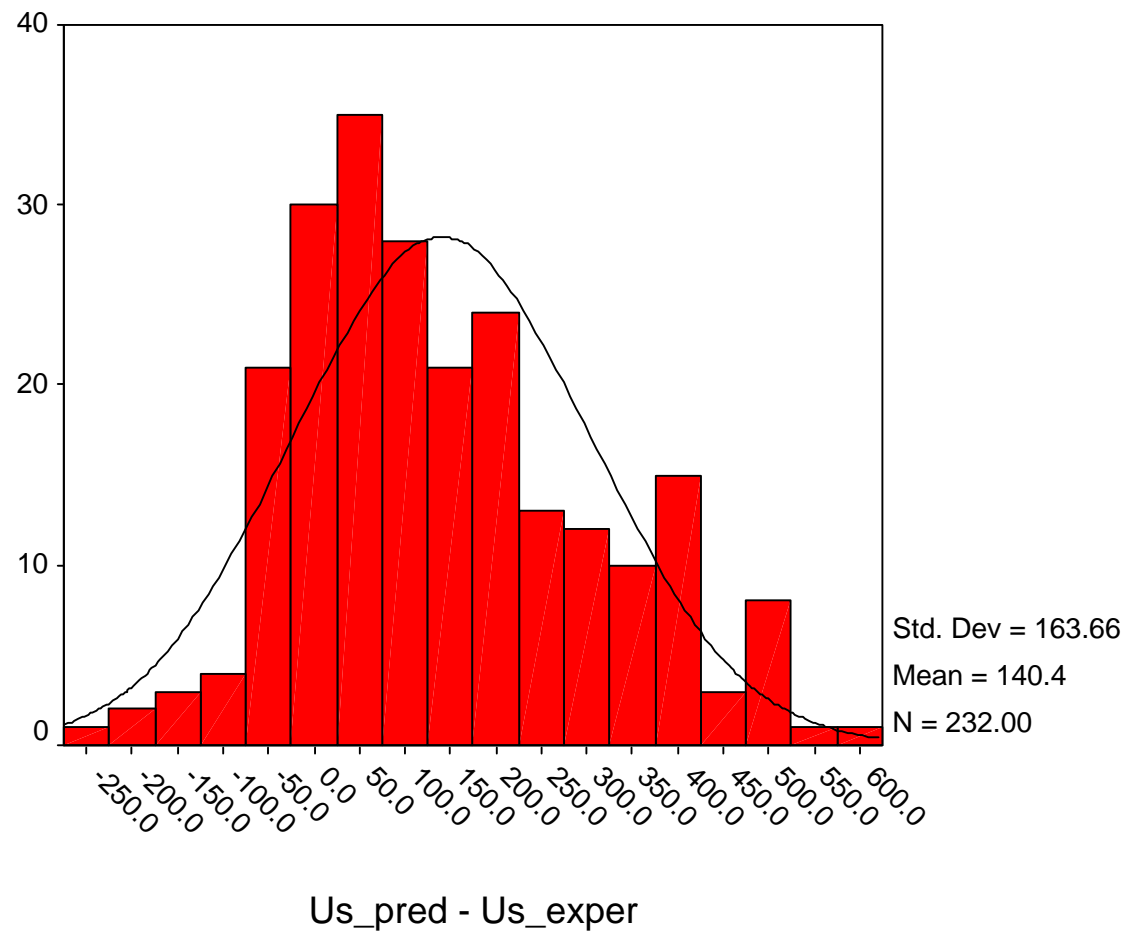
- Question 1: Is there sufficient evidence that the mean of the prediction error is not zero given the uncertainty reflected in the validation data? This is the simplest statistical question we can ask.
- Question 2: What can we say about the ability of the model to predict U_s and a function of U_p ?
- Question 3: Can we put error bounds on the prediction of U_s as a function of U_p ?

Statistical Inference

- Need model for prediction error uncertainty
- Initially assume prediction errors are independent and normally distributed
- Test normality

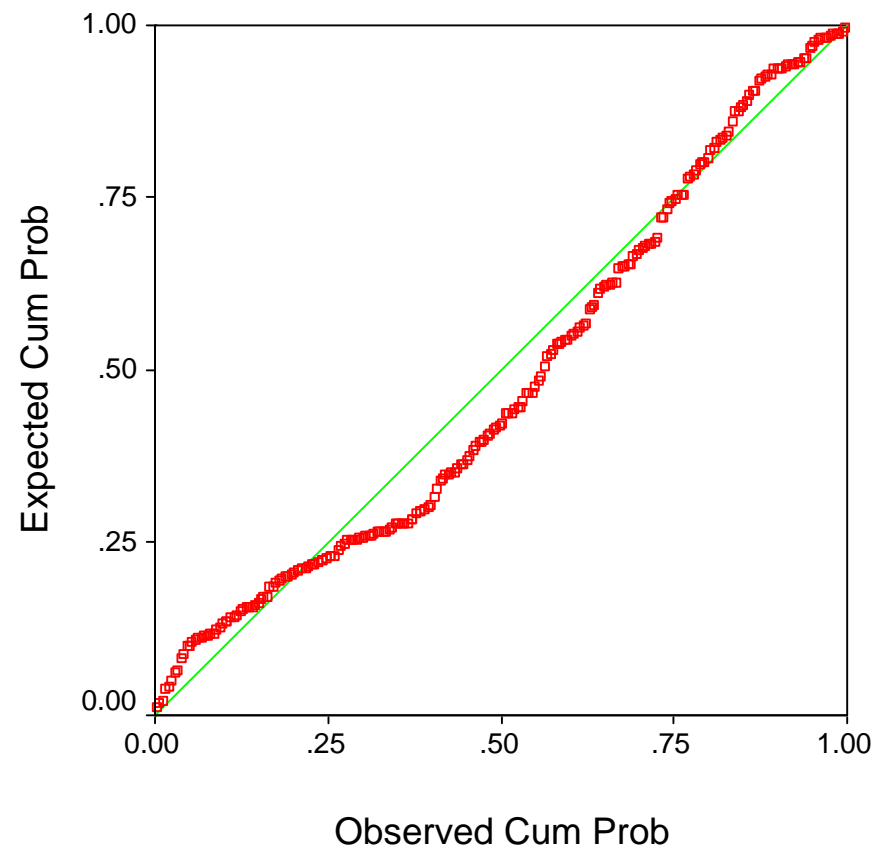


Distribution of Prediction Errors



Probability-Probability Plot

Normal P-P Plot of Us_pred - Us_exper



Test for Normality

One-Sample Kolmogorov-Smirnov Test Us_pred - Us_exper

N		232
Normal Parameters ^{a,b}	Mean	140.4
	Std. Deviation	163.7
Most Extreme Differences	Absolute	0.095
	Positive	0.095
	Negative	-0.059
Kolmogorov-Smirnov Z		1.452
Asymp. Sig. (2-tailed)		0.030

a. Test distribution is Normal.

b. Calculated from data.

Test for Normality, continued

- The level of significance is only 3% indicating that if the distribution of the prediction errors is normally distributed, we would have only a 3% chance of obtaining a maximum difference this far from normal
- We reject the hypothesis that the prediction errors are normally distributed at this level of significance
- This implies that we must use statistical methods for non-normally distributed prediction errors to test our model
- Tried other distributions, will use nonparametric methods

Nonparametric Methods

- Nonparametric methods do not require that the probability distribution be well characterized.
- Nonparametric methods are less powerful than parametric methods in that they are more likely to accept a bad model.
- These methods do require some assumptions such as independence of data.

The Sign Test

- Assume model valid if there are as many positive differences as negative differences for prediction errors
- Count the number of positive differences and evaluate the probability of this many positive differences assuming a symmetric distribution

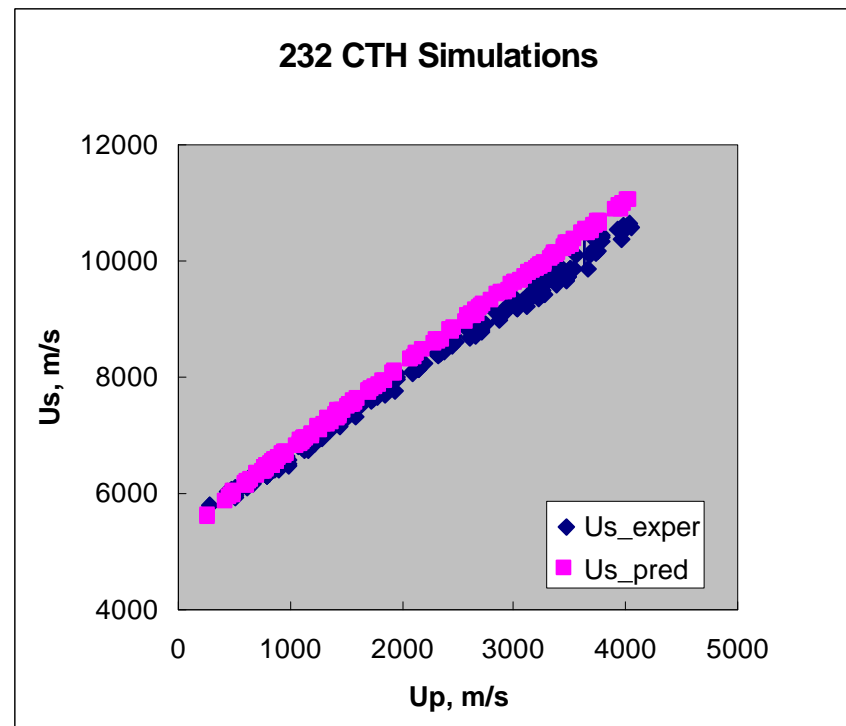
		N
Us_pred - Us_exper	Negative Differences	42
	Positive Differences	190
	Ties	0
	Total	232
Test statistic $Z = 9.651$		
Asymp. Significance (2-tailed) = 0.000		

Nonparametric Results

- The probability that a good model with this non-symmetry in the prediction error count is less than 1 in 1000
- Thus there is no significant evidence that this model has prediction errors symmetrically distributed about zero
- What can we say about the models ability to predict U_s as a function of U_p ?

What About $U_s = \text{Function}(U_p)$?

- U_{s_exper} and U_{s_pred} appear to both be linear in U_p .
- Perform linear regression and compare the regression coefficients



Regression

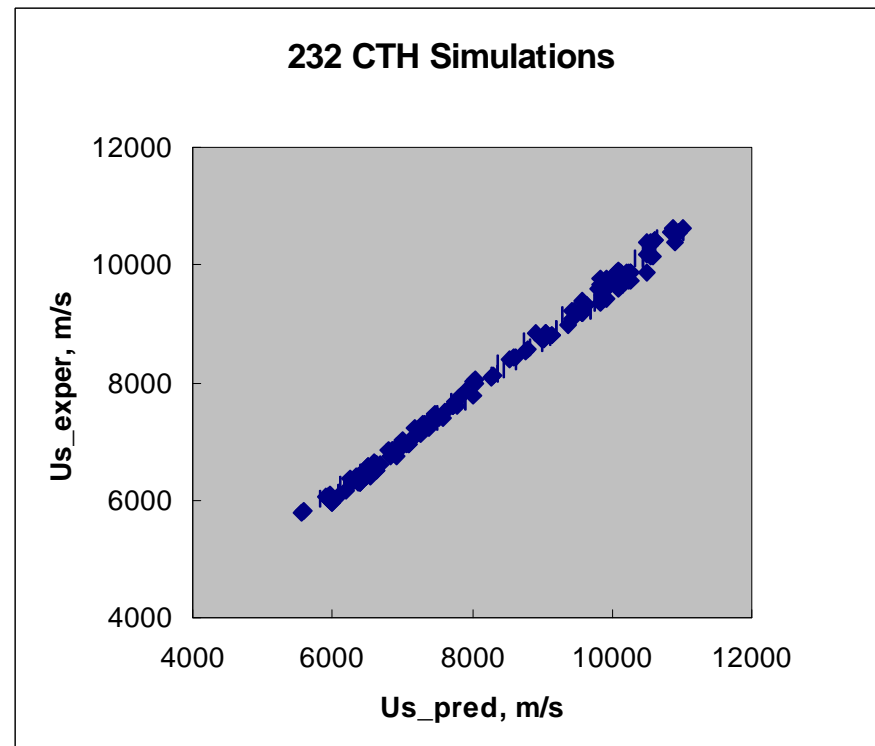
- $U_s = s U_p + \text{constant}$

Model		Coefficient	95% Confidence Interval for Coefficients	
			Lower Bound	Upper Bound
Exper:	Constant	5360.854	5340.903	5380.804
	s	1.299	1.290	1.308
Pred:	Constant	5256.690	5249.502	5263.878
	s	1.425	1.422	1.428

- Coefficients do not overlap
- Will use slightly different approach to investigate this further

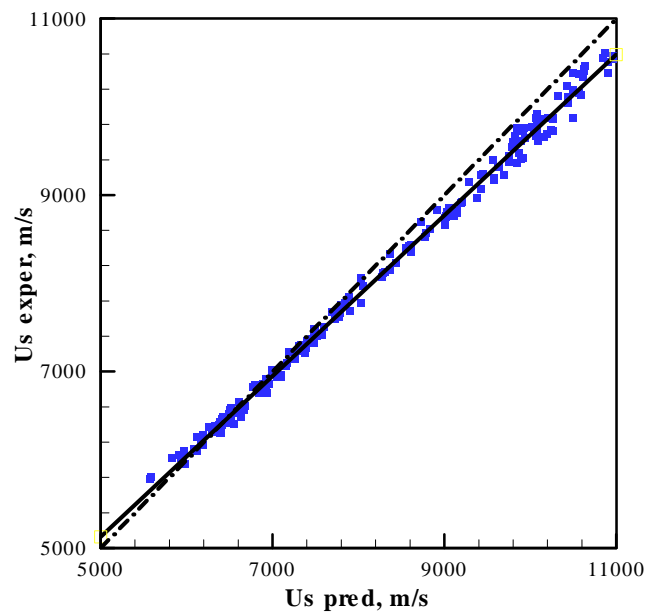
More General Approach

- We can plot Us_{exper} vs. Us_{pred}
- Perform regression on the results
- Evaluate whether slope=1 and intercept=0
- Statistical assumptions
 - no variability in Us_{pred}
 - errors in Us_{exper} are independent, normally distributed, with uniform variance, and uncorrelated

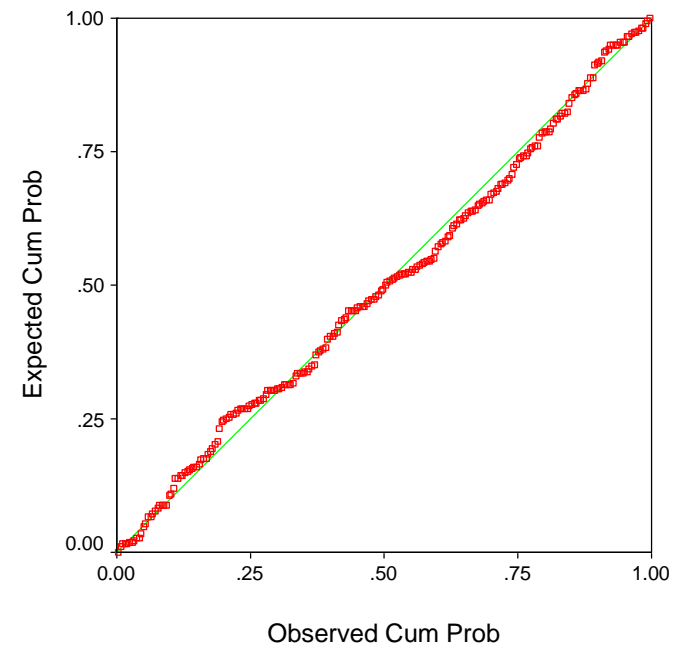


Us_exper vs. Us_pred

Regression



P-P Plot for Residuals



Test for Normality, Regression

One-Sample Kolmogorov-Smirnov Test Regression Residuals

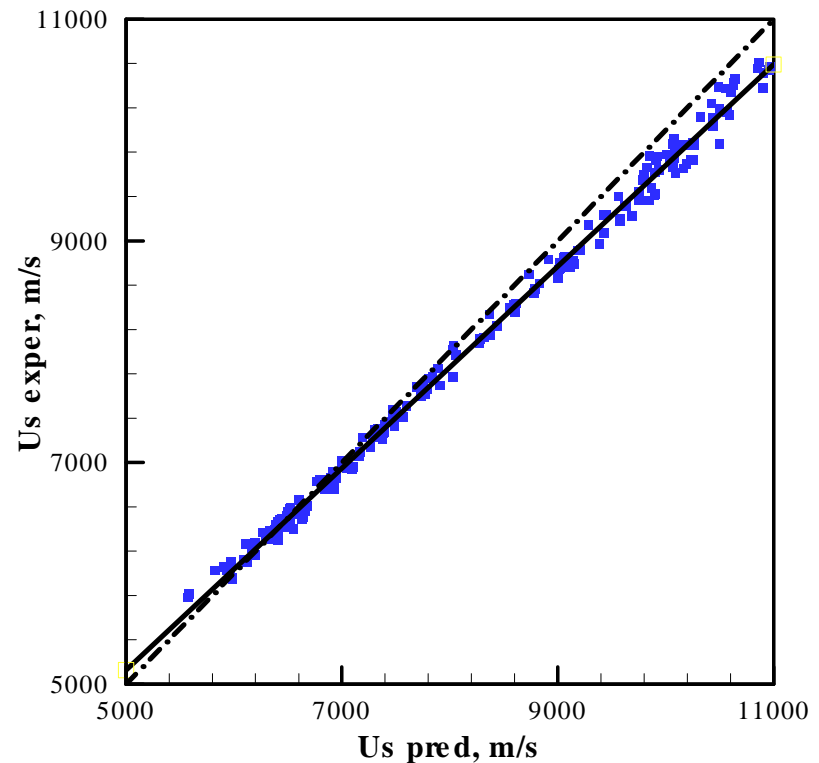
N		232
Normal Parameters ^{a,b}	Mean	0.000
	Std. Deviation	83.69
Most Extreme Differences	Absolute	0.050
	Positive	0.045
	Negative	-0.050
Kolmogorov-Smirnov Z		0.769
Asymp. Sig. (2-tailed)		0.596

a. Test distribution is Normal.

b. Calculated from data.

Regression of Experiment vs. Prediction

- Good evidence that prediction errors are normally distributed
- Can use normal distribution for statistical inference
- Is slope = 1?
- Is intercept = 0?



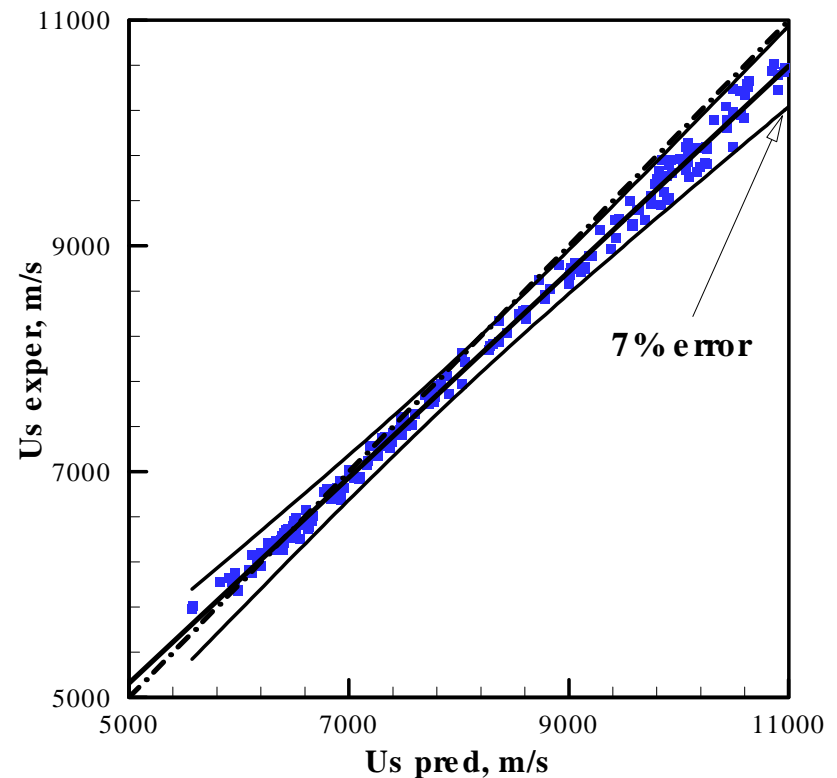
Regression of Us_exper vs. Us_pred

		95% Confidence Interval for Coefficients	
	Coefficient	Lower Bound	Upper Bound
intercept	570.492	514.473	626.510
slope	.911	.905	.918

- The intercept is not zero and the slope is not unity within their 95% confidence intervals - CTH model does not appear to be scientifically valid for this particular application

Prediction Bounds of U_{s_exper} vs. U_{s_pred}

- Thin curves - 95% prediction bounds on regression
- Use to define engineering bounds
- $< 5\%$ chance that a U_{s_exper} will lie outside the error bounds
- $< 5\%$ chance that U_{s_exper} will be different from U_{s_pred} by more than 7% at $U_{s_pred}=11000$



This Validation Exercise

- There is no sufficient statistical evidence that the model over predicts and under predicts in equal frequency based on the uncertainty of the validation experiment. Therefore, there is no evidence that the model is valid.
- Error bounds could be established on predictions of U_s as a function of U_p because we had sufficient data with easy to model statistical characteristics

This is as Easy as it Gets

- The data has very little scatter and there is sufficient repetition of the data
- The data is univariant and apparently independent
- The difference between the experiment and prediction is so large relative to estimated uncertainty that strong statements on model validity could be made using nonparametric methods
- A linear relation exists between the predicted and measured U_s , and the regression residuals were normally distributed, allowing one to easily establish a model for the bounds on the prediction error

Validation: Multivariate Data

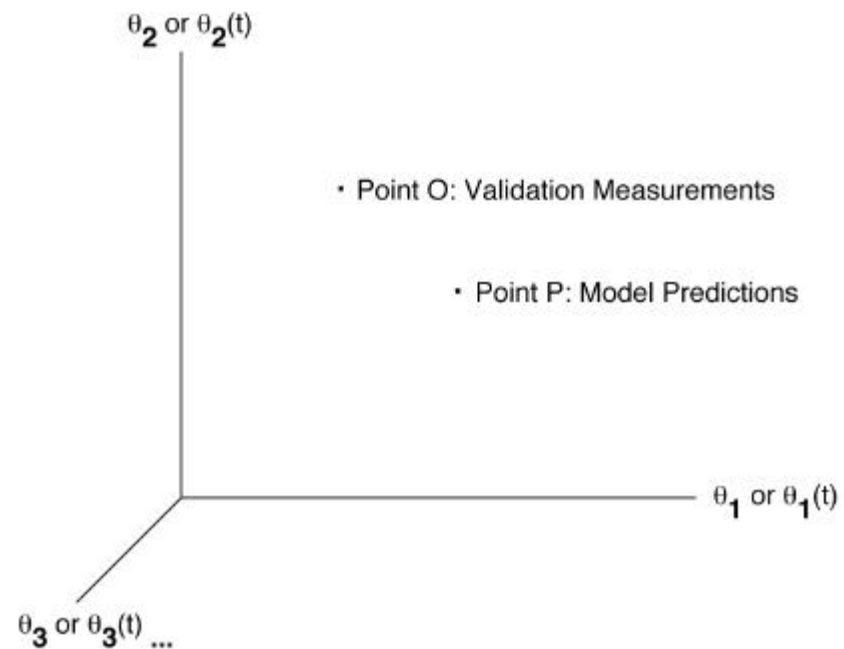
- There are many experimental situations for which it is simply not practical to perform multiple independent experiments
- Multivariate data is usually available, but the data are often highly correlated with an unknown correlation structure
- Because of the above, it is difficult to estimate the probability distributions using the data directly

Multivariate Data, continued

- If we have sufficient data to statistically characterize the model parameters and the measurement error, then we can use the model itself to develop a *model* for the probability distributions for the prediction errors (i.e., through Monte Carlo or sensitivity analysis).
- We can perform statistical model validation if the probability structure (including the correlation structure) can be identified

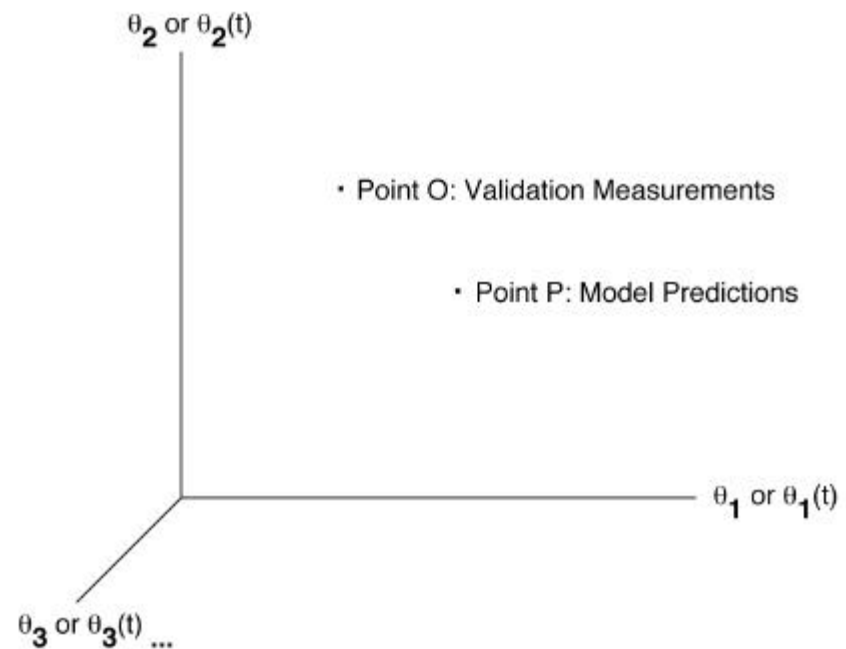
Observation/Prediction Space (n-dimensional)

- A geometric approach is useful to conceptualize validation with correlated errors
- Points O and P each represent n-tuples where
 - n is either the total number of measurements, or
 - n is the number of measurement locations at time t.



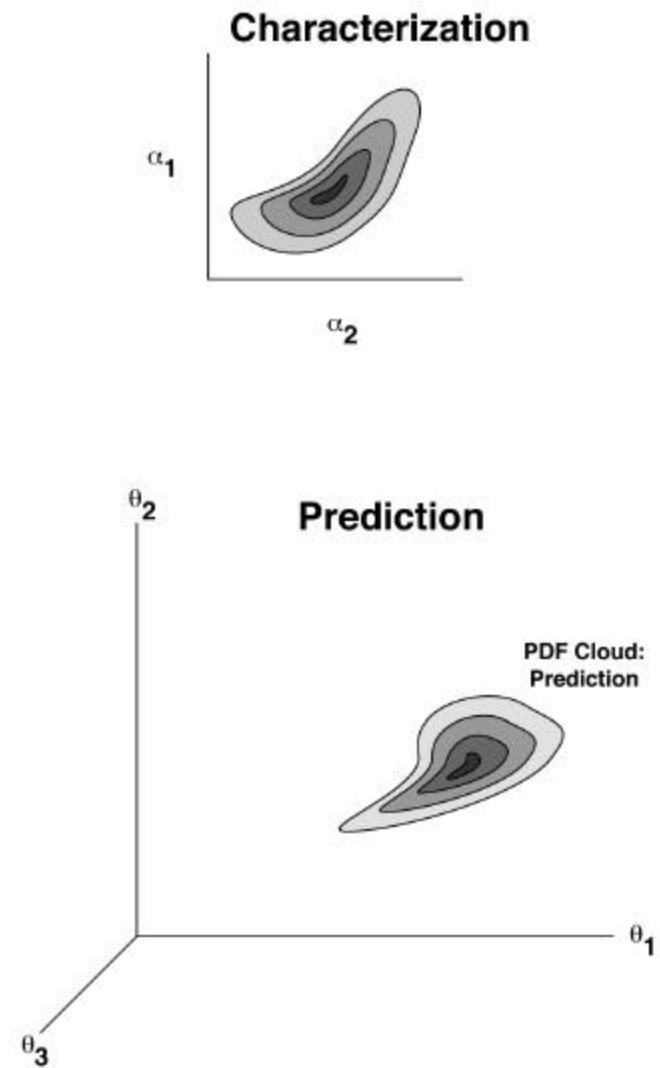
Model Validation

- Are these two points far enough apart to consider the model invalid?
- Are these two points far enough apart relative to the *modeled* uncertainty in the validation exercise to consider the model invalid?



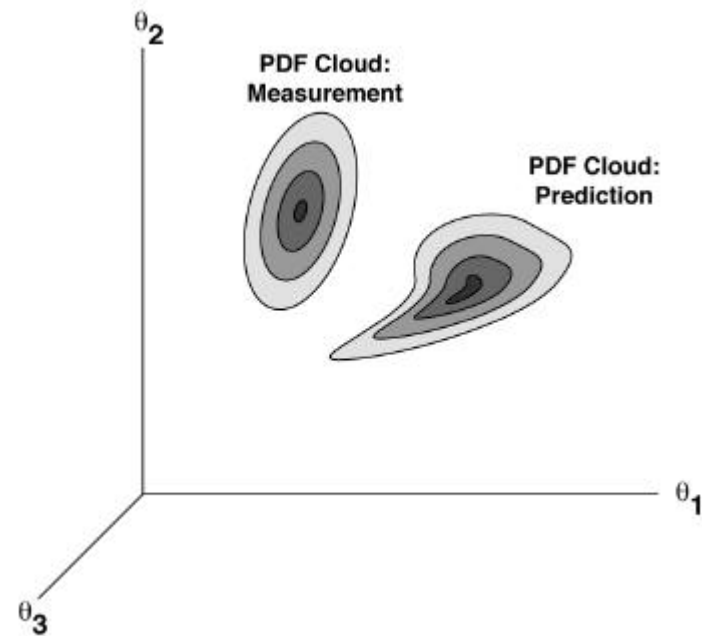
Uncertainty Analysis

- Characterize the model parameters and their uncertainty (variability)
- Propagate this uncertainty through the model to evaluate uncertainty (variability) in predictions due to parameter uncertainty (variability)



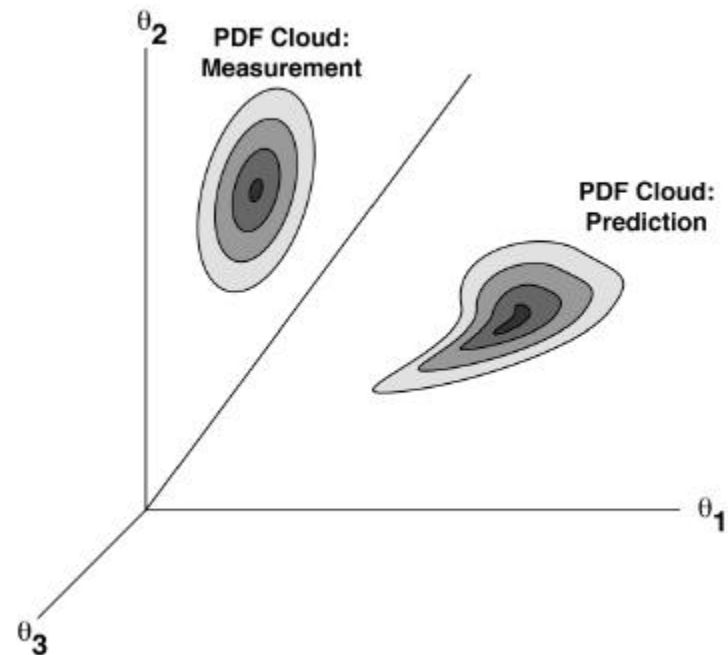
Validation

- Add measurement and its estimated uncertainty
- Is this model valid?
- Point Validation
 - Validation measures are in terms of the prediction errors directly
- Integrated Measure Validation
 - Validation measures are functionals of the prediction errors



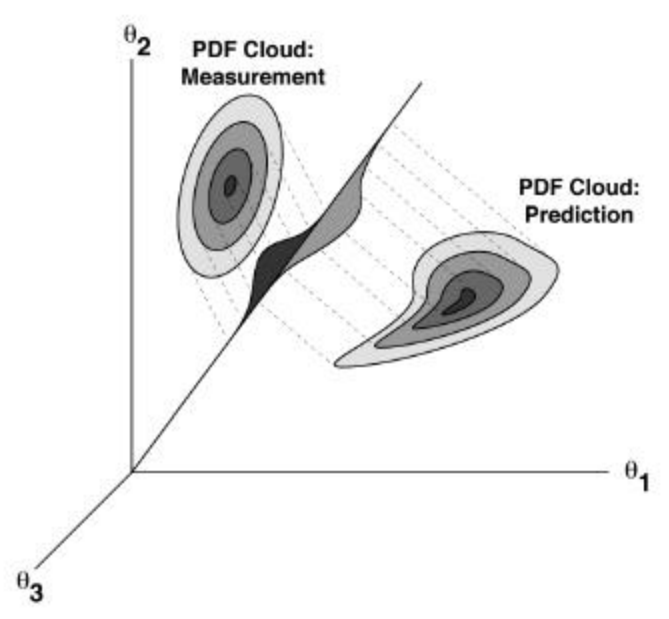
Integrated Measures

- Means
- Linear regression coefficients
- Mass balance over a region
- Integrated measures can often be represented as projections onto subspaces of the validation space



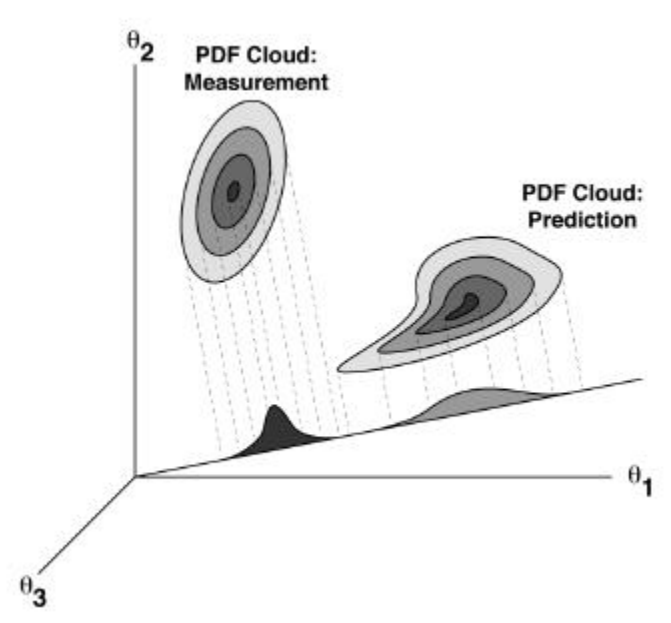
Projection onto Measure 1

- Is the distance between the prediction and measurement, as projected onto the validation measure, large relative to the uncertainty?
- Relative distance not large for this case - implies no significant evidence to reject the model



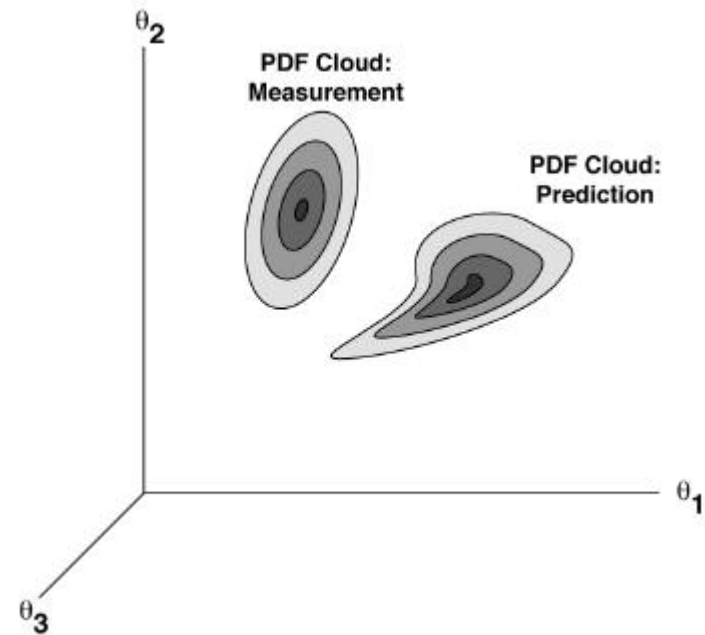
Projection onto Measure 2

- Is the distance between the prediction and measurement, as projected onto the validation measure, large relative to the uncertainty?
- Relative distance is large for this case - implies significant evidence to reject the model
- A model that appears valid by one integrated measure may appear invalid by another
- Integrated measure should be appropriate for the final application



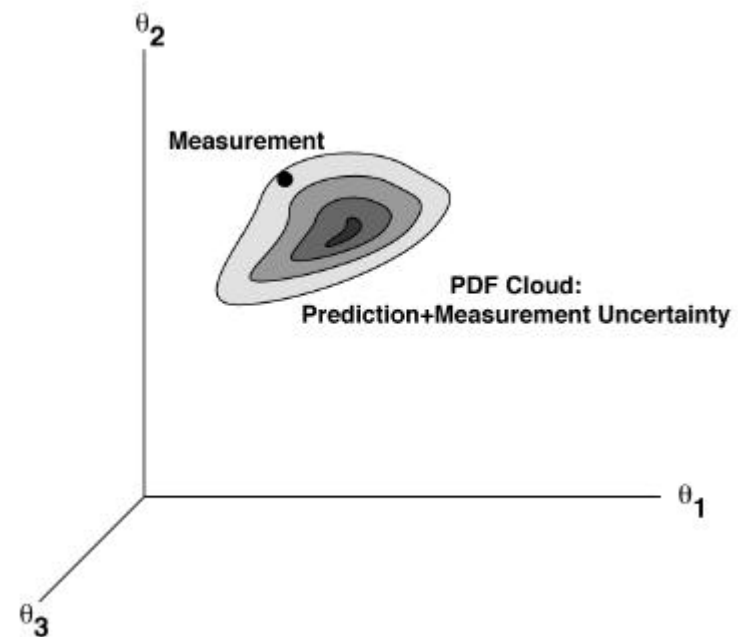
Point Validation

- Measure distance between prediction and measurement in n-dimensional space directly rather than in a subspace
- Is this distance large relative to the uncertainty on the validation exercise?
- Yes for this case, but not clear how to actually quantify this



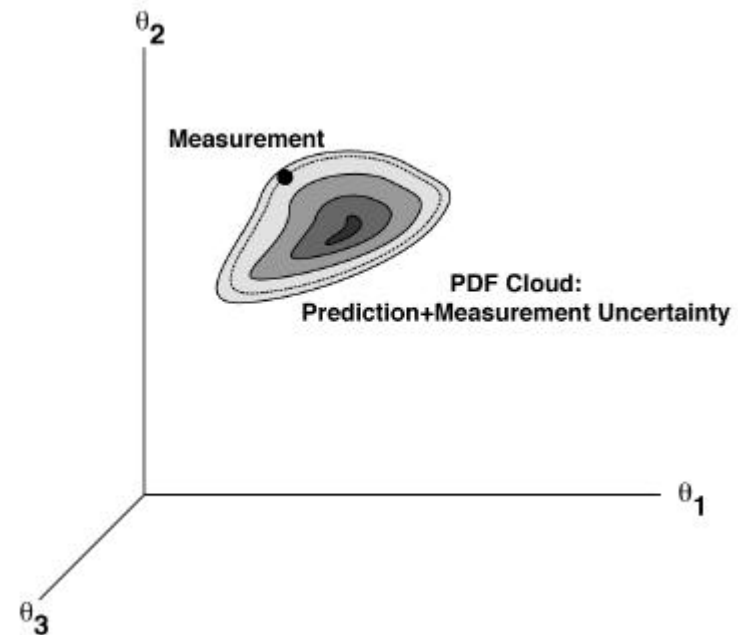
Point Validation, continued

- Easier to visualize if we combine the modeled uncertainty of the validation exercise into one uncertainty
 - Prediction uncertainty
 - Experimental uncertainty
- Is the distance between the measurement and the center of the PDF cloud large relative to the size of the cloud?
- If so, reject the model as valid
- How do we actually measure this distance?



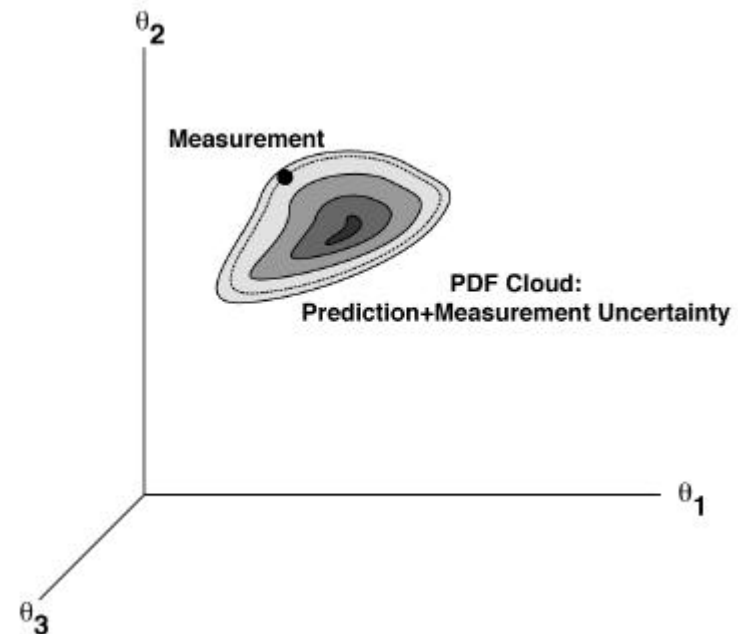
Possible Metric

- Metric: Define constant distance curves to be constant PDF curves
- Critical Distance: Reject the model if measurement is outside a critical constant PDF curve
- Analogous to two-tailed critical region for the univariate cases for symmetric distributions



Problems

- Metric arbitrarily defined relative to final application if validation experiment does not closely represent the final application
- Constant PDF curves may not be appropriate for application
- Very expensive to compute for n large



Application Defined Metrics

- One possibility to reduce the computational requirements is to use integrated metrics (i.e., metrics based on projections into lower dimensional spaces)
- Many applications have go/no-go decisions based on a small set of conditions (i.e., low degrees of freedom)
- The validation experiments usually measure different variables than those used for the go/no-go decisions
- Develop a map between the validation experiment variables and the decision variables using the model
- Use the map to define integrated validation measures

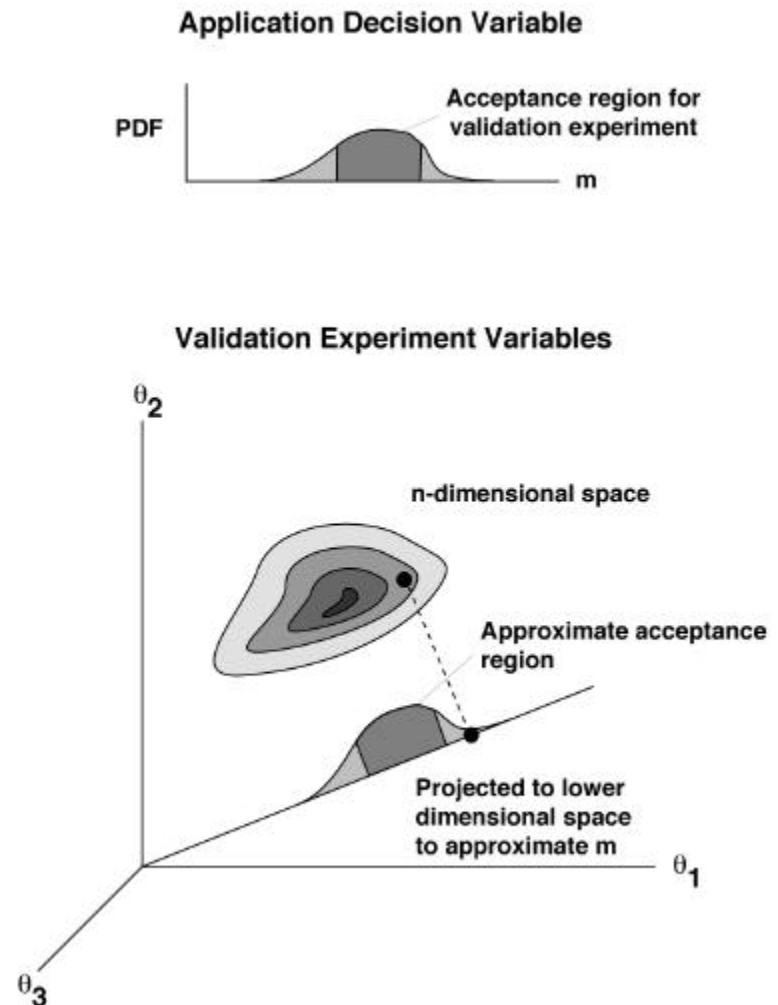
Example using Sensitivity Analysis

- m is the critical scalar measure for the application
- We wish to test model's ability to accurately predict m
- Sensitivity analysis used to approximate measure in validation space

$$\left. \begin{array}{l} \overline{\Delta \mathbf{q}} = \mathbf{A} \overline{\Delta \mathbf{a}} \\ \Delta m = \overline{b}^T \overline{\Delta \mathbf{a}} \end{array} \right\} \Rightarrow \Delta m \approx \overline{b}^T \mathbf{A}^+ \overline{\Delta \mathbf{q}}$$

\mathbf{A}^+ - pseudoinverse of \mathbf{A}

- Perform statistical inference on predicted m equal to experimental m in validation space



Application Defined Metric

- The characteristics of the pseudo-inverse of \mathbf{A} can be used to evaluate how well the application metric maps to the corresponding metric for the validation experiment (geophysical inverse theory literature)
- Poor mappings indicate that validation experiments not well suited for the final application
- These ideas may be extendable for application to multiple validation experiments

Summary

- The availability of sufficient independent experimental data to adequately characterize the probability distributions for the prediction error uncertainty greatly simplifies model validation - standard statistical methods can be used
- Correlated data adds significant complexity
 - Typical of measurements made over space and time
 - Difficult to evaluate correlation structure from data only
- Propagate the uncertainty to develop a model for the prediction uncertainty and correlation
 - Requires statistical characterization of model parameters and experimental data
 - Can be very CPU intensive for multivariate validation data due to high dimensionality in the validation space

Summary, continued

- Integrated measures reduce the CPU requirements due to their lower dimensionality
- Models that appear valid by one integrated measure can appear invalid by another
- Application specific integrated measures may be the most appropriate measured because they relate application metrics to validation metrics